Published online 27 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D295–D305***
doi:10.1093/nar/gks1229

dbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications

Cheng-Tsung Lu[1], Kai-Yao Huang[1], Min-Gang Su[1], Tzong-Yi Lee[1,2,*], Neil Arvin Bretaña[1], Wen-Chi Chang[3], Yi-Ju Chen[4], Yu-Ju Chen[4] and Hsien-Da Huang[5,6,*]

[1]Department of Computer Science and Engineering, Yuan Ze University, [2]Gradulate Program in Biomedical Informatics, Yuan Ze University, Chung-Li 320, [3]Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, [4]Institute of Chemistry, Academia Sinica, Taipei 115, [5]Institute of Bioinformatics and Systems Biology and [6]Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan

Received September 18, 2012; Revised October 26, 2012; Accepted October 31, 2012

ABSTRACT

Protein modification is an extremely important post-translational regulation that adjusts the physical and chemical properties, conformation, stability and activity of a protein; thus altering protein function. Due to the high throughput of mass spectrometry (MS)-based methods in identifying site-specific post-translational modifications (PTMs), dbPTM (http://dbPTM.mbc.nctu.edu.tw/) is updated to integrate experimental PTMs obtained from public resources as well as manually curated MS/MS peptides associated with PTMs from research articles. Version 3.0 of dbPTM aims to be an informative resource for investigating the substrate specificity of PTM sites and functional association of PTMs between substrates and their interacting proteins. In order to investigate the substrate specificity for modification sites, a newly developed statistical method has been applied to identify the significant substrate motifs for each type of PTMs containing sufficient experimental data. According to the data statistics in dbPTM, >60% of PTM sites are located in the functional domains of proteins. It is known that most PTMs can create binding sites for specific protein-interaction domains that work together for cellular function. Thus, this update integrates protein–protein interaction and domain–domain interaction to determine the functional association of PTM sites located in protein-interacting domains. Additionally, the information of structural topologies on transmembrane (TM) proteins is integrated in dbPTM in order to delineate the structural correlation between the reported PTM sites and TM topologies. To facilitate the investigation of PTMs on TM proteins, the PTM substrate sites and the structural topology are graphically represented. Also, literature information related to PTMs, orthologous conservations and substrate motifs of PTMs are also provided in the resource. Finally, this version features an improved web interface to facilitate convenient access to the resource.

INTRODUCTION

Protein post-translational modification (PTM) plays an essential role in various cellular processes that adjusts the physical and chemical properties, folding, conformation, stability and activity of proteins; thus altering protein function (1). More than 200 different types of PTMs have been identified by mass spectrometry (MS)-based proteomics (2). The biological functions of this ubiquitous regulatory mechanisms include phosphorylation for signal transduction, attachment of fatty acids for membrane anchoring and association, glycosylation for changing protein half-life, targeting substrates, promotion of cell–cell and cell–matrix interactions, acetylation and methylation of histone for gene regulation and ubiquitylation for protein degradation (3). With the

*To whom correspondence should be addressed. Tel: +886 3 4638800 (ext. 3007); Fax: +886 3 4638850; Email: francis@saturn.yzu.edu.tw
Correspondence may also be addressed to Hsien-Da Huang. Tel: +886 3 5712121 (ext. 56952); Fax: +886 3 5739320; Email:bryan@mail.nctu.edu.tw

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

high-throughput MS or MS/MS-based methods in proteomics, several databases associated with a specific modification type have been established. Phospho.ELM (4), Phosphorylation Site Database (5), PhosphoSitePlus (6), PHOSIDA (7) and PhosPhAt (8) were developed for accumulating experimentally verified phosphorylation sites. NetworKIN (9) and RegPhos (10) designed an integrative method to identify the kinase-substrate phosphorylation networks. O-GLYCBASE (11) and dbOGAP (12) are the databases of glycoproteins, most of which include experimentally verified O-linked glycosylation sites. UbiProt (13) stores experimental ubiquitylated proteins and ubiquitylation sites, which are implicated in protein degradation through an intracellular ATP-dependent proteolytic system. PupDB (14) is a prokaryotic ubiquitin-like protein (Pup) database which stores a collection of experimentally identified pupylated proteins and pupylation sites from published articles. It also integrates the information of pupylated proteins with corresponding structures and functional annotations. An increasing number of proteomic studies have suggested that protein *S*-nitrosylation plays important role in the nitric oxide (NO)-related redox pathway. With this, a new database named dbSNO (15) was established by manually curating *S*-nitrosylation peptides from research articles.

With regard to public resources of multiple PTM types currently available, UniProtKB/Swiss-Prot (2,16) includes as much information of PTMs as is available with functional and structural annotations. SysPTM (17) has designed a systematic platform for multi-type PTM research and data mining. Additionally, Human Protein Reference Database (HPRD) (18) contains a wealth of information relevant to the function of human proteins in health and disease, as well as the annotation of PTMs. With the importance of protein modifications in biological processes, we have previously proposed dbPTM (19) which integrates published databases in order to obtain experimentally validated protein modifications, as well as putative PTM substrate sites predicted by a series of accurate computational tools (20–22). Version 2.0 of dbPTM was extended to a knowledge base comprising the modified sites, solvent accessibility of substrate, protein secondary and tertiary structures, protein domains and protein variations (23).

Due to the high throughput of MS/MS-based methods in identifying site-specific PTMs, this version (dbPTM 3.0) not only integrates experimental PTMs from public resources but also manually curates MS/MS peptides associated with PTMs from research articles using a text mining approach. The dbPTM 3.0 aims to be an informative resource for investigating the substrate specificity of PTM sites and functional association of PTMs between substrates and their interacting proteins. In order to investigate the substrate specificity for modification sites, a newly developed method, MDDLogo (24), has been applied to identify the significant substrate motifs for each type of PTMs. According to the data statistics in dbPTM, >60% of PTM sites are located in protein functional domains. Many PTMs can create binding sites for specific protein-interaction domains that work together for cellular function and read the state of proteome to cellular organization (25). Thus, this update integrates both protein–protein interaction (PPI) and domain–domain interaction information to determine the functional association of PTM sites located in protein-interacting domains. Additionally, in order to delineate the structural correlation between the reported PTM sites and transmembrane (TM) topologies, the information of structural topologies on TM proteins is integrated in dbPTM 3.0. To facilitate the investigation of PTMs on TM proteins, PTM sites as well as the structural topology of TM proteins are graphically represented. Furthermore, the web interface is enhanced to facilitate access to the resource and is now freely accessible at http://dbPTM.mbc.nctu.edu.tw/.

IMPROVEMENTS

The highlighted improvements and advances in dbPTM 3.0 are presented in Figure 1 including data integration from public PTM resources and research articles, investigation of PTM substrate site specificity, investigation of PTM-associated protein interactions, as well as the investigation of the effects of PTM on TM proteins. To facilitate the study of PTMs and their functions, the web interface is redesigned and enhanced. Published literature information related to PTMs, orthologous conservations and substrate motifs of PTM sites are also provided in this online resource. The details of each improved process are depicted as follows.

Data integration from public PTM resources and research articles

Supplementary Figure S1 shows the detailed system flow of the construction of dbPTM 3.0. Due to the inaccessibility of database contents in several online PTM resources, a total 11 biological databases related to PTMs are integrated in dbPTM, including UniProtKB/Swiss-Prot (2), version 9.0 of Phospho.ELM (4), PhosphoSitePlus (6), PHOSIDA (26), version 6.0 of O-GLYCBASE (11), dbOGAP (12), dbSNO (15), version 1.0 of UbiProt (13), PupDB (14), version 1.1 of SysPTM (17) and release 9.0 of HPRD (27). A brief description and the data statistics of the integrated databases are given in Supplementary Table S1. To solve the heterogeneity among the data collected from different sources, the reported modification sites are mapped to the UniProtKB protein entries using sequence comparison. With the high throughput of MS-based methods in post-translational proteomics, this update also includes manually curated MS/MS-identified peptides associated with PTMs from research articles through a literature survey. First, a table list of PTM-related keywords is constructed by referring to the UniProtKB/SwissProt PTM list (http://www.uniprot.org/docs/ptmlist.txt) and the annotations of RESID (28). Then, all fields in the PubMed database are searched based on the keywords of the constructed table list. This is then followed by downloading the full text of the research articles. For the various experiments of proteomic identification, a text-mining system is developed to survey full-text literature that

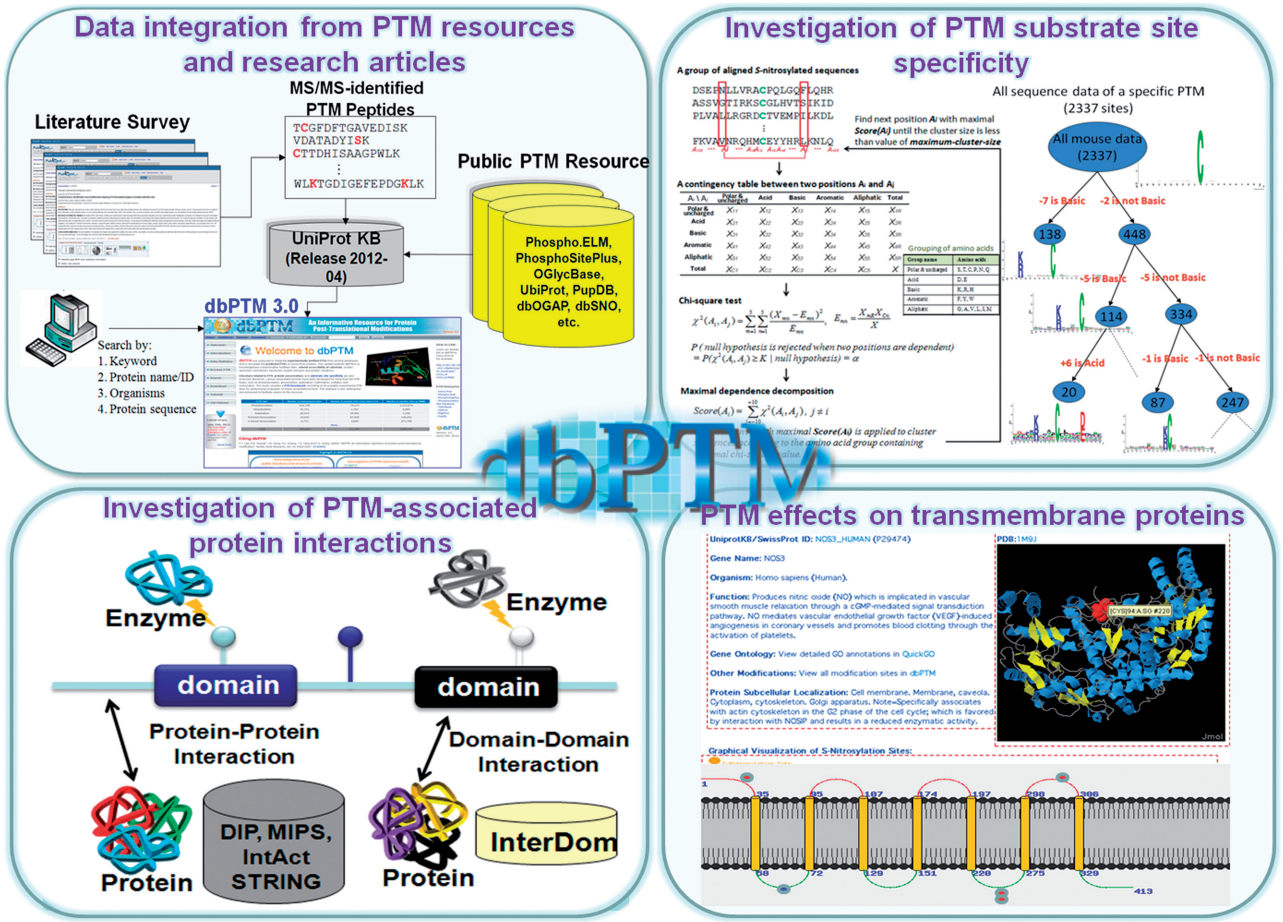

Figure 1. The highlighted improvements and advances in dbPTM 3.0.

potentially describes the site-specific identification of modified sites. Approximately 800 original and review articles associated with MS/MS proteomics and protein modifications are retrieved from PubMed (July 2012). Next, the full-length articles are manually reviewed for precisely extracting the MS/MS peptides along with the modified sites. Furthermore, in order to determine the locations of PTMs on a full-length protein sequence, the experimentally verified MS/MS peptides are then mapped to UniProtKB protein entries based on its database identifier (ID) and sequence identity. In the process of data mapping, MS/MS peptides that cannot align exactly to a protein sequence are discarded. Finally, each mapped PTM site is attributed with a corresponding literature (PubMed ID).

Detection of PTM substrate site specificities

Due to the difficulty of detecting the conserved motifs for a specific PTM with a large data size, MDDLogo (24) was used to identify the substrate motifs for each type of PTMs containing >500 modified peptides. MDDLogo exploits maximal dependence decomposition (MDD) in order to discover conserved motifs from a group of aligned signal sequences. MDD groups a set of aligned signal sequences into subgroups that capture the most significant dependencies between positions. MDD adopts Chi-squared test $\chi^2(A_i,A_j)$ to evaluate the dependence of amino acid occurrence between two positions A_i and A_j that surround the PTM substrate sites. MDDLogo has demonstrated its effectiveness in identifying substrate motifs of plant and virus phosphorylation (29,30), as well as the mouse *S*-nitrosylation (31). In order to extract the motifs that have conserved biochemical property of amino acids when doing MDD, it categorizes the 20 types of amino acids into five groups such as aliphatic, polar and uncharged, acid, basic and aromatic groups, as shown in Supplementary Figure S2. An example of MDD clustering on *S*-nitrosylation data shows that position −7 has the maximal dependence with the occurrence of basic amino acids, including lysine (k), arginine (r) and histidine (H). Subsequently, all data can be divided into two subgroups: one has the occurrence of basic amino acids in position −7 and the other does not have the occurrence of basic amino acids in position −7. The MDD clustering is a recursive process to divide the data sets into tree-like subgroups.

Integration of protein domains, domain–domain interactions and PPIs

Protein-interaction domains usually recognize short peptide motifs of a target protein but do not bind stably

until the peptides have the appropriate PTMs; this can create binding sites for specific protein-interaction domains that work together for cellular function and read the state of proteome to cellular organization (25). For instance, the SH2 domain can bind to phosphotyrosine (pTyr)-associated peptides in a manner that depends on ligand phosphorylation and the motif of the flanking amino acids (32,33). Thus, this update integrates the information of protein functional domains and PPIs to infer the PTM-dependent protein interactions. To investigate the preference of functional domains for PTM, this study refers to the domain annotations in InterPro (34). InterPro is an integrated resource, which was developed initially as a means of rationalizing the complementary efforts of the PROSITE (35), PRINTS (36), Pfam (37) and ProDom (38) databases, for providing protein 'signatures' such as protein families, domains and functional sites. For the information of experimentally verified PPIs, five databases including DIP (39), MINT (40), IntAct (41), HPRD (18) and STRING (42) are integrated in dbPTM (see Supplementary Table S2). Additionally, the domain–domain interactions of InterDom (43) are also integrated to determine the functional association for the PTM sites which locate in protein-interacting domains.

Integration of TM proteins with structural topology

TM proteins play crucial roles in various cellular processes (44). A genome-wide study has discovered that ~20–30% of the proteins encoded by a typical genome are TM proteins (45). However, due to the experimental difficulties in obtaining high-quality structures, TM proteins are notably under-represented in Protein Data Bank (46). The biological roles of PTMs playing on TM proteins include phosphorylation for signal transduction and ion transport, acetylation for structure stability, attachment of fatty acids for membrane anchoring and association, as well as the glycosylation for receptors targeting, cell–cell interactions and virus infection (44,47). With the importance of PTMs functioning on TM proteins, the experimentally curated information of membrane topologies is collected from TMPad (48), TOPDB (49), PDB_TM (50) and OPM (51). In order to provide a comprehensive investigation of TM proteins, a potential set of TM proteins is extracted from UniProtKB (52) by choosing protein entries which contain the keyword 'TRANSMEM' in feature ('FT') line, the localization of 'membrane' and the information of TM topology. The potential TM proteins are further filtered using a TM prediction program MEMSAT (53) to determine its membrane topologies. As shown in Supplementary Table S3, the filtering process resulted in 2216 experimental and 43 142 potential TM proteins with membrane topologies. To facilitate the investigation of PTMs on TM proteins, the structural topology of TM proteins is graphically represented using PHP GD library, as well as the PTM substrate sites. Moreover, the tertiary structures of TM proteins and PTM sites are visualized using the Jmol program (54).

Integration of external biological databases

For a given protein, the basic biological functions can be obtained from the annotations of UniProtKB. To provide more information about protein functional and structural annotations relevant to the modified proteins and the PTM substrate sites, the data contents of Gene Ontology (GO) (55), Protein Data Bank (PDB) (46) and Clusters of Orthologous Groups (COGs) (56) have been integrated in dbPTM. In this study, the information regarding the molecular function, cellular components and biological process for a modified protein can be accessed by a crosslink that refers to the corresponding entry from QuickGO (57) via a UniProtKB accession number. In order to facilitate the investigation of structural characteristics surrounding the PTM substrate sites, protein tertiary structure obtained from PDB was graphically presented by Jmol program. For proteins with tertiary structures (5% of UniProtKB/Swiss-Prot proteins), the protein structural properties, such as solvent accessibility and secondary structure of residues, were calculated by DSSP (58). With respect to the previous studies investigating the structural characteristics of PTMs (59–61) in proteins without known tertiary structures, two effective tools, RVP-net (62) and PSIPRED (63), are used to predict the solvent accessibility and secondary structure, respectively. In order to observe whether a PTM sites located in the conserved regions among orthologous protein sequences, the COGs of proteins were integrated and the ClustalW (64) program was adopted to implement the alignment of multiple protein sequences in each COG cluster.

DATA CONTENT AND UTILITY

Data statistics of the integrated PTM sites

In order to provide the most comprehensive data of PTMs, this update not only integrates experimental PTMs from 11 external PTM-related resources but also manually curates MS/MS peptides associated with PTMs from ~800 research articles. After removing the redundancy data among these heterogeneous resources, there are totally 208 521 experimental PTM sites in dbPTM 3.0. All the experimental PTM sites are further categorized by PTM types and the number of non-redundant PTM sites is calculated. As the data statistics of representative PTM types shown in Table 1, protein phosphorylation contains the most abundant data of experimentally verified substrate sites. Due to the high throughput of Ms/MS-based proteomics in the site-specific identification of modified peptides, several PTMs have a significantly increasing number of experimental data, including protein ubiquitylation, acetylation, methylation, N-linked and O-linked glycosylation, as well as the emerging *S*-nitrosylation. In addition to the experimental PTM sites, UniProtKB/Swiss-Prot provides putative PTM sites by using sequence similarity or evolutionary potential, which are annotated as 'by similarity', 'potential' or 'probable' in the 'MOD_RES' fields. A total of 226 122 putative sites for all PTM types are integrated in dbPTM.

Table 1. Data statistics of experimental and putative PTM sites in dbPTM

PTM types	Number of experimental substrate sites	Number of putative substrate sites from UniProtKB/Swiss-Prot	Number of HMM-predicted sites
Phoshorylation	142 446	74 174	1 414 879
Ubiquitylation	23 647	1702	8865
N-linked glycosylation	15 242	87 529	418 253
Acetylation	9683	19 981	1156
O-linked glycosylation	3508	3695	373 758
Amidation	2533	1445	114 034
Hydroxylation	1629	1274	9743
Methylation	1585	5479	22 332
Pyrrolidone carboxylic acid	829	742	12 322
Sumoylation	725	800	13 042
Gamma-carboxyglutamic acid	448	814	1942
Palmitoylation	312	5252	33 830
Sulfation	207	800	70 005
Myristoylation	178	1275	988
C-linked glycosylation	156	99	3923
Prenylation	130	1327	6741
Nitration	80	93	1432
Deamidation	52	165	2022
S-nitrosylation	3096	170	–
Oxidation	333	180	–
ADP-ribosylation	140	164	–
N6-succinyllysine	88	69	–
Formylation	56	125	–
GPI anchoring	34	849	–
Bromination	33	56	–
N6-malonyllysine	33	167	–
Citrullination	32	110	–
N6-carboxylysine	30	1566	–
Glutathionylation	19	32	–
FAD	19	163	–
Others	1218	15 825	–
Total	208 521	226 122	2 509 267

Moreover, a KinasePhos-like method (19–22) has been adopted to construct the profile hidden Markov models (HMMs) for 18 types of PTM. Especially in protein phosphorylation, >70 kinase-specific prediction models are constructed and used to identify the putative phosphorylation sites with their kinases. These models were applied to search the potential PTM sites against UniProtKB/Swiss-Prot protein sequences. As given in Table 1, totally 2 509 267 putative sites for all PTM types are detected by HMMs with 90% predictive specificity. All the experimental PTM sites and putative PTM sites are available and downloadable in the web interface.

Enhanced web interface

To facilitate the use of the dbPTM resource, the web interface has been redesigned and enhanced to allow efficient access to the protein of interest. Supplementary Figure S3 shows the content of a typical dbPTM query: (i) quick search by IDs and keywords, (ii) basic information, (iii) graphical visualization of PTM sites with structural characteristics and functional domains, (iv) table of experimental PTM sites with reported literature, (v) orthologous conservation of PTM substrate sites, (vi) PPIs and domain–domain interactions and (vii) literature related to PTMs. The combined visualization of PTM sites and function domains for a protein sequence can help users to understand the functional associations of PTM substrate sites. According to the multiple sequence alignment result of orthologous proteins, users can investigate whether a PTM site located in evolutionary conserved regions, which indicates that the orthologous sites in other species could be involved in the same modification. Additionally, this update incorporates the protein functional domains and domain–domain interactions to infer the PTM-dependent protein interactions. Moreover, the literatures associated with PTMs are categorized by the modification type.

In addition to the database query by the protein name, gene name, UniProtKB ID or accession, the protein sequence is allowed for homology search against UniProtKB protein sequence database using Blast (65) program. For browse function of dbPTM web site, a summary table of PTM types and their modified residues is provided for users to efficiently access the number of data in a specific modified amino acid of a PTM type. The annotations of PTM types are referred to the UniProtKB/Swiss-Prot PTM list (http://www.uniprot.org/docs/ptmlist.txt). As depicted in Supplementary Figure S4, the acetylation of lysine (K) is chosen to obtain more detailed information such as the location of the modification in protein sequence, the modified chemical formula, the mass difference and the substrate site specificity, which is the preference of amino acids surrounding the modification sites. The structural

characteristics, such as solvent accessibility and secondary structure surrounding the PTM substrate sites, are also provided. Additionally, the substrate site specificity of the acetylated lysines is investigated in detail with reference to the subcellular localizations of acetylated proteins. Previous work has demonstrated that the co-localization of acetyltransferases and substrate proteins could be a promising method to investigate the substrate site specificities and could be adopted to improve the computational identification of protein acetylation sites (66).

Investigation of PTM substrate site specificities

Given a window length, n, the fragment of $2n+1$ residues centering on PTM site (position 0) is extracted and the positional frequencies of amino acids are calculated and presented as sequence logos by WebLogo (67). Supplementary Figure S5 shows the substrate motif and structural characteristics of experimental phosphorylation sites. According to the kinase classification extracted from KinBase (http://kinase.com/) and RegPhos (10), the substrate site specificity of protein phosphorylation could be further categorized into >200 kinase groups. As given in Supplementary Figure S5, most of the kinase-specific substrate motifs have conserved amino acids surrounding the phosphorylation sites. For the PTMs other than phosphorylation, there are no annotations of catalytic enzymes or transferases due to the experimental difficulty in identifying the catalytic enzymes for a specific PTM. Based on the basic concept of sequence conservation, a sequence logo could display the substrate motif for each PTM type with a group of aligned sequences. However, it is difficult to explore conserved motifs for large-scale sequence data; for instance, a sequence logo for all phosphorylation data involved with various catalytic kinases fails to obviously present the kinase-specific substrate specificity. Thus, for the PTM containing sufficient data of experimental substrate sites, MDDLogo was performed to cluster a group of aligned substrate sequences into subgroups containing statistically significant motifs. As the example of protein *S*-nitrosylation presented in Figure 2, 10 sequence logos, which were identified from 3095 *S*-nitrosylated peptides with a 13-mer window length,

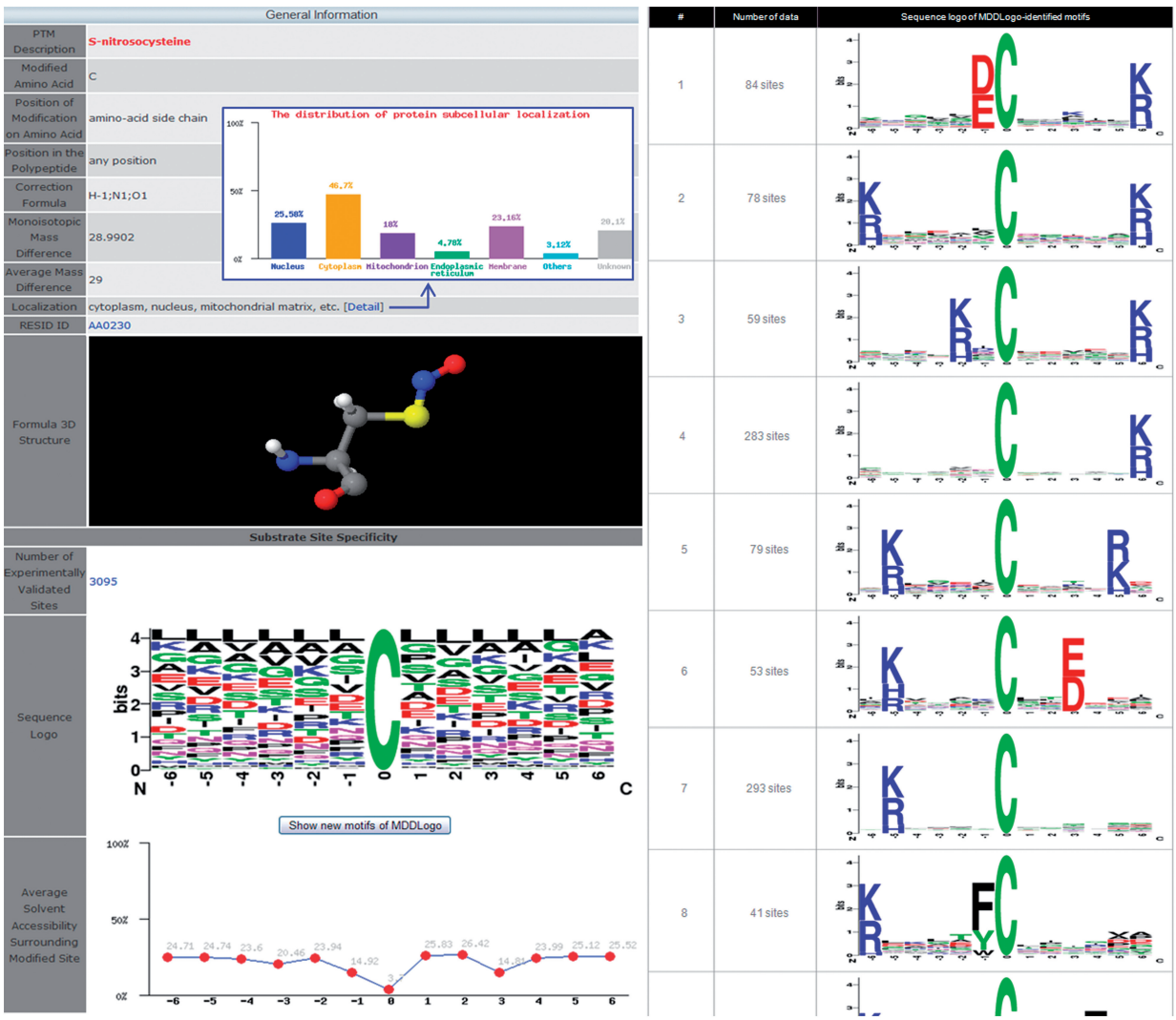

Figure 2. The MDDLogo-identified substrate motifs of protein S-nitrosylation sites.

contain a conserved motif of positively charged amino acids (K, R and H) surrounding the *S*-nitrosocysteine. Interestingly, the first and sixth groups contain the conserved motifs of negatively charged amino acids (D and E) accompanied by positively charged amino acids at two specific positions. Consistent with previous studies (68–73), the *S*-nitrosylated cysteines may be located within an acid-base motif flanked by acidic and basic amino acids.

Investigation of PTM-associated domains and protein interactions

According to the data statistics in dbPTM, >60% of experimentally verified PTM sites locate in the functional domains of proteins. Such statistics could be analyzed in detail for each type of PTMs. For instance of protein *S*-nitrosylation, which is an emerging PTM playing crucial role in the regulation of NO-related cellular processes, the statistics shows that ~70% of the reported *S*-nitrosylation sites locate within the functional domains. Furthermore, the detailed distribution of functional domains covering *S*-nitrosylation sites is given in Supplementary Table S4. It is observed that the most preferred functional domain is the 'nucleotide-binding alpha–beta plait' with InterPro ID: IPR012677 which covers 47 *S*-nitrosylation sites. Another preferred functional domain is the 'RNA recognition motif, RNP-1' domain with InterPro ID: IPR000504 which covers 46 *S*-nitrosylation sites. This investigation indicates that these *S*-nitrosylation sites may play important roles in the domains of proteins involving in DNA or RNA binding (74). In addition, Supplementary Table S5 shows the distribution of functional domains covering substrate sites for several representative PTMs, including acetylation, methylation, hydroxylation, N-linked and O-linked glycosylation, phosphorylation and ubiquitylation.

Many PTMs provide binding sites for specific protein-interaction domains, which often contain a conserved structure for the modified site and a more flexible surface for the flanking amino acids, synergize to regulate cellular processes (75–78). In order to investigate the PTM-associated protein interactions, the information of domain–domain interactions collected from InterDom is adopted in this study. As the case study of 'Histone H3' (UniProtKB ID: H31_HUMAN) presented in Figure 3, 'Heterochromatin protein 1 homolog alpha' ('HP1', UniProtKB ID: CBX5_HUMAN) and 'WD repeat-containing protein 5' ('WDR5', UniProtKB ID: WDR5_HUMAN) interact with 'Histone H3'. When

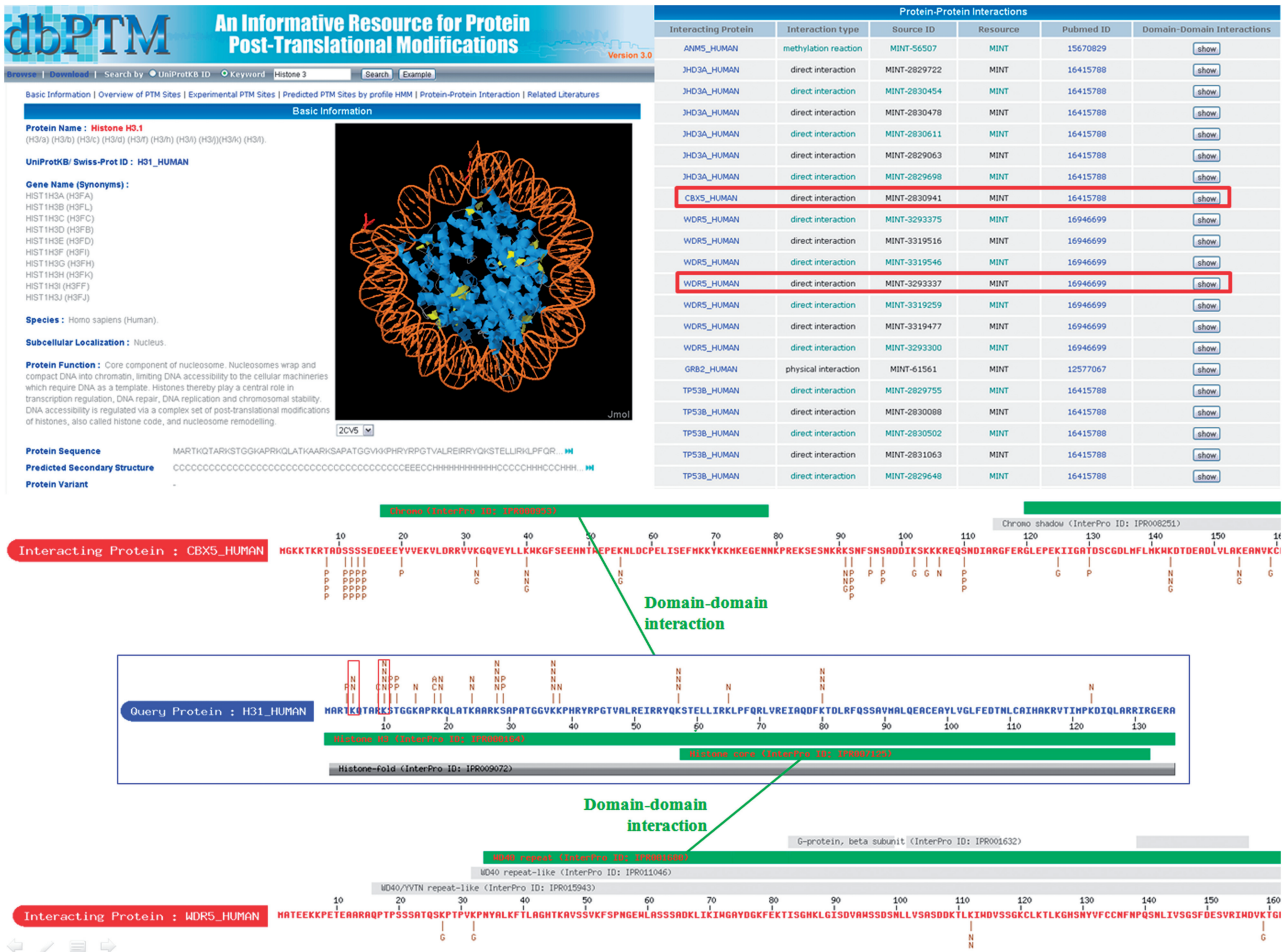

Figure 3. A case study of domain–domain interactions and PTM-associated protein interactions on Histone H3 (UniProtKB ID: H31_HUMAN).

investigating the protein interaction between 'HP1' and 'Histone H3' in detail, there is a domain–domain interaction between 'Chromodomain' (InterPro ID: IPR000953) and 'Histone H3' (InterPro ID: IPR000164). Among the PTMs located in the domain of 'Histone H3', a previous study has demonstrated that the 'HP1 chromodomain' can bind to the 'Histone H3' methylated at lysine 10 (79). Another protein interaction shows that there is a domain–domain interaction between the 'WD40 Repeat' (InterPro ID: IPR001680) and 'Histone Core' (InterPro ID: IPR007125). It has been proposed that the structural motif for the specific recognition of methylated 'Histone H3' lysine 5 by 'WD40 Repeat' of 'WDR5' is essential to vertebrate development (80,81). This investigation indicates that the other PTM sites could be the potential binding sites for protein-interaction domains.

Investigation of PTM sites on TM proteins

According to the data statistics of PTM sites and TM proteins in dbPTM, a total of 9644 and 68 775 PTM substrate sites locate on the 2088 experimental and 33 747 potential TM proteins, respectively. In order to investigate the structural distribution of PTM sites on TM proteins, the structural topologies of a TM protein are mainly categorized into four types: extracellular, cytoplasmic, TM and unknown regions. Supplementary Table S6 provides the structural distribution of PTMs containing >10 substrate sites on experimental TM proteins. Interestingly, without the consideration of substrate sites located in unknown region, all of the N-linked (GlcNAc...) glycosylation sites are located in the extracellular region, as well as the O-linked and C-linked glycosylation sites. This investigation is reasonable to understand the biological effect of glycosylation functioning on TM proteins for receptor targeting and cell–cell interactions (47). Otherwise, the phosphorylation sites are mainly located in cytoplasmic regions, which induce signal transduction and ion transport. The structural distribution of PTM sites could be the means to infer the potential roles of PTMs functioning on TM proteins. Actually, a previous work has demonstrated that the incorporation of membrane topology could improve the performance of predicting O-linked glycosylation sites on TM proteins (82). Supplementary Figure S6 shows a graphical visualization of the PTMs and membrane topology on human Beta-2 adrenergic receptor (ADRB2). Furthermore, two modification sites Tyr141 (pTyr) and Cys341 (*S*-palmitoyl cysteine) are further highlighted in red on the tertiary structure (PDB ID: 2R4R) using Jmol viewer,

Table 2. Advances and improvements in this update (dbPTM 3.0)

Features	dbPTM 1.0	dbPTM 2.0	dbPTM 3.0
Protein entry	UniProtKB/Swiss-Prot (release 46)	UniProtKB/Swiss-Prot (release 55)	UniProtKB release 2012-04
Experimental PTM resource	UniProtKB/Swiss-Prot, Phospho.ELM and O-GLYCBASE	UniProtKB/Swiss-Prot, Phospho.ELM, PHOSIDA, HPRD, O-GLYCBASE and UbiProt	UniProtKB/Swiss-Prot, HPRD, SysPTM, Phospho.ELM, PhosphoSitePlus, PHOSIDA, O-GLYCBASE, dbOGAP, dbSNO, UbiProt and PupDB
Literature survey of PTMs	–	–	>5000 modified peptides extracted from ~800 articles
Literatures related to PTMS	–	Yes	Yes (categorized by PTM types)
Computationally predicted PTMs	Phosphorylation, glycosylation and sulfation	20 types of PTM	18 types of PTM
Protein tertiary structure	Protein Data Bank (PDB)	Protein Data Bank (PDB)	Protein Data Bank (PDB)
Structural properties of PTM sites	Amino acid frequency	Amino acid frequency, solvent accessibility and secondary structure	Amino acid frequency, solvent accessibility, secondary structure and intrinsic disorder region
PTM annotation	RESID (373 PTM annotations)	RESID (431 PTM annotations)	RESID (431 PTM annotations)
Kinase family annotation	–	KinBase	KinBase and RegPhos
Protein functional domain	InterPro	InterPro	InterPro and InterProScan
Protein–protein interaction	–	–	DIP, MINT, IntAct, HPRD and STRING
Domain–domain interaction	–	–	InterDom
Functional association of PTM	–	–	PTM-associated domains and PTM-dependent protein interactions
PTM substrate motif	–	WebLogo	WebLogo and MDDLogo
Evolutionary conservation of PTM sites	–	ClustalW	ClustalW and COG
Transmembrane topology	–	–	TMPad, PDBTM, TOPDB and OPM
Graphical visualization	PTM, solvent accessibility, protein variation and protein domain	PTM, solvent accessibility, secondary structure, protein variation, protein domain, tertiary structure, orthologous conservation and sequence logo	PTM, solvent accessibility, secondary structure, protein variation, protein domain, tertiary structure, orthologous conservation, sequence logo, PTM substrate motifs, domain–domain interaction, protein–protein interaction, transmembrane topology and tertiary structure of PTMs

which indicates the solvent accessibility and distance between them.

CONCLUSION

The expansion of the dbPTM database increases its usefulness for researchers investigating the impact of PTMs on protein function and cellular processes. Additionally, the enhanced web interface enables both wet-lab biologists and bioinformatics researchers to efficiently explore the further information about protein PTMs. Table 2 summarizes the advancements and new features supported in dbPTM 3.0. In the future, we expect dbPTM to continue to grow with the increasing availability of data in resources such as Phospho.ELM, PhosphoSitePlus and UniProtKB. One area that we can envision dbPTM improving greatly in prospective works is implementing a more accurate method for the discovery of PTM substrate motifs. Also, enhancements on the text mining algorithm will enable the system to select MS/MS peptides from research articles associated with protein modifications with a higher confidence rate. In order to provide more adequate information for PTM function, the descriptions associated with the biological function of PTMs will be extracted from research articles using an information retrieval system. Moreover, the thermodynamic parameters for proteins (83), PPIs (84) and protein–nucleic acid interactions (85) could be integrated for the investigation of PTM-associated protein stability.

AVAILABILITY

The data content of dbPTM will be regularly maintained and semiannually updated. The resource is now available at http://dbPTM.mbc.nctu.edu.tw/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6 and Supplementary Figures 1–6.

FUNDING

National Science Council of the Republic of China financial support, [contract no. 101-2628-E-155-002-MY2, NSC 101-2311-B-009-003-MY3, NSC 100-2627-B-009-002, NSC 101-2911-I-009-101 and NSC 101-2319-B-400-001]. Funding for open access charge: National Science Council of Taiwan.

Conflict of interest statement. None declared.

REFERENCES

1. Mann,M. and Jensen,O.N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, **21**, 255–261.
2. Farriol-Mathis,N., Garavelli,J.S., Boeckmann,B., Duvaud,S., Gasteiger,E., Gateau,A., Veuthey,A.L. and Bairoch,A. (2004) Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics*, **4**, 1537–1550.
3. Seo,J. and Lee,K.J. (2004) Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J. Biochem. Mol. Biol.*, **37**, 35–44.
4. Dinkel,H., Chica,C., Via,A., Gould,C.M., Jensen,L.J., Gibson,T.J. and Diella,F. (2011) Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res.*, **39**, D261–D267.
5. Wurgler-Murphy,S.M., King,D.M. and Kennelly,P.J. (2004) The Phosphorylation Site Database: a guide to the serine-, threonine-, and/or tyrosine-phosphorylated proteins in prokaryotic organisms. *Proteomics*, **4**, 1562–1570.
6. Hornbeck,P.V., Kornhauser,J.M., Tkachev,S., Zhang,B., Skrzypek,E., Murray,B., Latham,V. and Sullivan,M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.
7. Gnad,F., Ren,S., Cox,J., Olsen,J.V., Macek,B., Oroshi,M. and Mann,M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, **8**, R250.
8. Heazlewood,J.L., Durek,P., Hummel,J., Selbig,J., Weckwerth,W., Walther,D. and Schulze,W.X. (2008) PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.*, **36**, D1015–D1021.
9. Linding,R., Jensen,L.J., Pasculescu,A., Olhovsky,M., Colwill,K., Bork,P., Yaffe,M.B. and Pawson,T. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.*, **36**, D695–D699.
10. Lee,T.Y., Bo-Kai Hsu,J., Chang,W.C. and Huang,H.D. (2011) RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans. *Nucleic Acids Res.*, **39**, D777–D787.
11. Gupta,R., Birch,H., Rapacki,K., Brunak,S. and Hansen,J.E. (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res.*, **27**, 370–372.
12. Wang,J., Torii,M., Liu,H., Hart,G.W. and Hu,Z.Z. (2011) dbOGAP—an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinformatics*, **12**, 91.
13. Chernorudskiy,A.L., Garcia,A., Eremin,E.V., Shorina,A.S., Kondratieva,E.V. and Gainullin,M.R. (2007) UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics*, **8**, 126.
14. Tung,C.W. (2012) PupDB: a database of pupylated proteins. *BMC Bioinformatics*, **13**, 40.
15. Lee,T.Y., Chen,Y.J., Lu,C.T., Ching,W.C., Teng,Y.C. and Huang,H.D. (2012) dbSNO: a database of cysteine S-nitrosylation. *Bioinformatics*, **28**, 2293–2295.
16. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
17. Li,H., Xing,X., Ding,G., Li,Q., Wang,C., Xie,L., Zeng,R. and Li,Y. (2009) SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol. Cell Proteomics*, **8**, 1839–1849.
18. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
19. Lee,T.Y., Huang,H.D., Hung,J.H., Huang,H.Y., Yang,Y.S. and Wang,T.H. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.*, **34**, D622–D627.
20. Huang,H.D., Lee,T.Y., Tzeng,S.W., Wu,L.C., Horng,J.T., Tsou,A.P. and Huang,K.T. (2005) Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J. Comput. Chem.*, **26**, 1032–1041.
21. Huang,H.D., Lee,T.Y., Tzeng,S.W. and Horng,J.T. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**, W226–W229.
22. Wong,Y.H., Lee,T.Y., Liang,H.K., Huang,C.M., Wang,T.Y., Yang,Y.H., Chu,C.H., Huang,H.D., Ko,M.T. and Hwang,J.K. (2007) KinasePhos 2.0: a web server for identifying protein

kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **35**, W588–W594.
23. Lee,T.Y., Hsu,J.B., Chang,W.C., Wang,T.Y., Hsu,P.C. and Huang,H.D. (2009) A comprehensive resource for integrating and displaying protein post-translational modifications. *BMC Res. Notes*, **2**, 111.
24. Lee,T.Y., Lin,Z.Q., Hsieh,S.J., Bretana,N.A. and Lu,C.T. (2011) Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics*, **27**, 1780–1787.
25. Seet,B.T., Dikic,I., Zhou,M.M. and Pawson,T. (2006) Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell Biol.*, **7**, 473–483.
26. Gnad,F., Gunawardena,J. and Mann,M. (2011) PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res.*, **39**, D253–D260.
27. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database—2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
28. Garavelli,J.S. (2004) The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics*, **4**, 1527–1533.
29. Lee,T.Y., Bretana,N.A. and Lu,C.T. (2011) PlantPhos: using maximal dependence decomposition to identify plant phosphorylation sites with substrate site specificity. *BMC Bioinformatics*, **12**, 261.
30. Bretana,N.A., Lu,C.T., Chiang,C.Y., Su,M.G., Huang,K.Y., Lee,T.Y. and Weng,S.L. (2012) Identifying protein phosphorylation sites with kinase substrate specificity on human viruses. *PLoS One*, **7**, e40694.
31. Lee,T.Y., Chen,Y.J., Lu,T.C. and Huang,H.D. (2011) SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. *PLoS One*, **6**, e21849.
32. Bradshaw,J.M. and Waksman,G. (2002) Molecular recognition by SH2 domains. *Adv. Protein Chem.*, **61**, 161–210.
33. Verkhivker,G.M., Bouzida,D., Gehlhaar,D.K., Rejto,P.A., Schaffer,L., Arthurs,S., Colson,A.B., Freer,S.T., Larson,V., Luty,B.A. *et al.* (2001) Hierarchy of simulation models in predicting molecular recognition mechanisms from the binding energy landscapes: structural analysis of the peptide complexes with SH2 domains. *Proteins*, **45**, 456–470.
34. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
35. Bairoch,A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.*, **19(Suppl.)**, 2241–2245.
36. Attwood,T.K., Beck,M.E., Bleasby,A.J. and Parry-Smith,D.J. (1994) PRINTS—a database of protein motif fingerprints. *Nucleic Acids Res.*, **22**, 3590–3596.
37. Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
38. Corpet,F., Gouzy,J. and Kahn,D. (1998) The ProDom database of protein domain families. *Nucleic Acids Res.*, **26**, 323–326.
39. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.
40. Chatr-Aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2006) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
41. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
42. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
43. Ng,S.K., Zhang,Z., Tan,S.H. and Lin,K. (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.*, **31**, 251–254.
44. Vinothkumar,K.R. and Henderson,R. (2010) Structures of membrane proteins. *Q. Rev. Biophys.*, **43**, 65–158.
45. Wallin,E. and von Heijne,G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.*, **7**, 1029–1038.
46. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
47. Ackers,G.K. and Smith,F.R. (1985) Effects of site-specific amino acid modification on protein interactions and biological function. *Annu. Rev. Biochem.*, **54**, 597–629.
48. Lo,A., Cheng,C.W., Chiu,Y.Y., Sung,T.Y. and Hsu,W.L. (2011) TMPad: an integrated structural database for helix-packing folds in transmembrane proteins. *Nucleic Acids Res.*, **39**, D347–D355.
49. Tusnady,G.E., Kalmar,L. and Simon,I. (2008) TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res.*, **36**, D234–D239.
50. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33**, D275–D278.
51. Lomize,M.A., Lomize,A.L., Pogozheva,I.D. and Mosberg,H.I. (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.
52. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
53. Nugent,T. and Jones,D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.
54. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.
55. Consortium,T.G.O. (2011) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
56. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
57. Binns,D., Dimmer,E., Huntley,R., Barrell,D., O'Donovan,C. and Apweiler,R. (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, **25**, 3045–3046.
58. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
59. Shien,D.M., Lee,T.Y., Chang,W.C., Hsu,J.B., Horng,J.T., Hsu,P.C., Wang,T.Y. and Huang,H.D. (2009) Incorporating structural characteristics for identification of protein methylation sites. *J. Comput. Chem.*, **30**, 1532–1543.
60. Lu,C.T., Chen,S.A., Bretana,N.A., Cheng,T.H. and Lee,T.Y. Carboxylator: incorporating solvent-accessible surface area for identifying protein carboxylation sites. *J. Comput. Aided Mol. Des.*, **25**, 987–995.
61. Lee,T.Y., Hsu,J.B., Lin,F.M., Chang,W.C., Hsu,P.C. and Huang,H.D. N-Ace: using solvent accessibility and physicochemical properties to identify protein N-acetylation sites. *J. Comput. Chem.*, **31**, 2759–2771.
62. Ahmad,S., Gromiha,M.M. and Sarai,A. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, **19**, 1849–1851.
63. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
64. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
65. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and

PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
66. Lee,T.Y., Hsu,J.B., Lin,F.M., Chang,W.C., Hsu,P.C. and Huang,H.D. (2010) N-Ace: using solvent accessibility and physicochemical properties to identify protein N-acetylation sites. *J. Comput. Chem.*, **31**, 2759–2771.
67. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
68. Hao,G., Derakhshan,B., Shi,L., Campagne,F. and Gross,S.S. (2006) SNOSID, a proteomic method for identification of cysteine S-nitrosylation sites in complex protein mixtures. *Proc. Natl Acad. Sci. USA*, **103**, 1012–1017.
69. Greco,T.M., Hodara,R., Parastatidis,I., Heijnen,H.F., Dennehy,M.K., Liebler,D.C. and Ischiropoulos,H. (2006) Identification of S-nitrosylation motifs by site-specific mapping of the S-nitrosocysteine proteome in human vascular smooth muscle cells. *Proc. Natl Acad. Sci. USA*, **103**, 7420–7425.
70. Lane,P., Hao,G. and Gross,S.S. (2001) S-nitrosylation is emerging as a specific and fundamental posttranslational protein modification: head-to-head comparison with O-phosphorylation. *Sci STKE*, **2001**, re1.
71. Stamler,J.S., Toone,E.J., Lipton,S.A. and Sucher,N.J. (1997) (S)NO signals: translocation, regulation, and a consensus motif. *Neuron*, **18**, 691–696.
72. Greco,T.M., Hodara,R., Parastatidis,I., Heijnen,H.G., Dennehy,M.K., Liebler,D.C. and Ischiropoulos,H. (2006) Identification of S-nitrosylation motifs by site-specific mapping of the S-nitrosocysteine proteome in human vascular smooth muscle cells. *Proc. Natl Acad. Sci. USA*, **103**, 7420–7425.
73. Chen,Y.-J., Ku,W.-C., Lin,P.-Y., Chou,H.-C., Khoo,K.-H. and Chen,Y.-J. (2010) S-alkylating labeling strategy for site-specific identification of the s-nitrosoproteome. *J. Proteome Res.*, **9**, 6417–6439.
74. delaTorre,A., Schroeder,R.A., Bartlett,S.T. and Kuo,P.C. (1998) Differential effects of nitric oxide-mediated S-nitrosylation on p50 and c-jun DNA binding. *Surgery*, **124**, 137–141; discussion 141–132.
75. Su,D., Hu,Q., Li,Q., Thompson,J.R., Cui,G., Fazly,A., Davies,B.A., Botuyan,M.V., Zhang,Z. and Mer,G. (2012) Structural basis for recognition of H3K56-acetylated histone H3-H4 by the chaperone Rtt106. *Nature*, **483**, 104–107.
76. Umehara,T., Nakamura,Y., Jang,M.K., Nakano,K., Tanaka,A., Ozato,K., Padmanabhan,B. and Yokoyama,S. (2010) Structural basis for acetylated histone H4 recognition by the human BRD2 bromodomain. *J. Biol. Chem.*, **285**, 7610–7618.
77. Owen,D.J., Ornaghi,P., Yang,J.C., Lowe,N., Evans,P.R., Ballario,P., Neuhaus,D., Filetici,P. and Travers,A.A. (2000) The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase gcn5p. *EMBO J.*, **19**, 6141–6149.
78. Durocher,D., Taylor,I.A., Sarbassova,D., Haire,L.F., Westcott,S.L., Jackson,S.P., Smerdon,S.J. and Yaffe,M.B. (2000) The molecular basis of FHA domain:phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms. *Mol. Cell*, **6**, 1169–1182.
79. Nielsen,P.R., Nietlispach,D., Mott,H.R., Callaghan,J., Bannister,A., Kouzarides,T., Murzin,A.G., Murzina,N.V. and Laue,E.D. (2002) Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature*, **416**, 103–107.
80. Wysocka,J., Swigut,T., Milne,T.A., Dou,Y., Zhang,X., Burlingame,A.L., Roeder,R.G., Brivanlou,A.H. and Allis,C.D. (2005) WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell*, **121**, 859–872.
81. Han,Z., Guo,L., Wang,H., Shen,Y., Deng,X.W. and Chai,J. (2006) Structural basis for the specific recognition of methylated histone H3 lysine 4 by the WD-40 protein WDR5. *Mol. Cell*, **22**, 137–144.
82. Chen,S.A., Lee,T.Y. and Ou,Y.Y. (2010) Incorporating significant amino acid pairs to identify O-linked glycosylation sites on transmembrane proteins and non-transmembrane proteins. *BMC Bioinformatics*, **11**, 536.
83. Gromiha,M.M., An,J., Kono,H., Oobatake,M., Uedaira,H. and Sarai,A. (1999) ProTherm: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Res.*, **27**, 286–288.
84. Kumar,M.D. and Gromiha,M.M. (2006) PINT: protein-protein Interactions Thermodynamic Database. *Nucleic Acids Res.*, **34**, D195–D198.
85. Prabakaran,P., An,J., Gromiha,M.M., Selvaraj,S., Uedaira,H., Kono,H. and Sarai,A. (2001) Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics*, **17**, 1027–1034.